DOCUMENT RESUME

ED 255 724 CE 041 234

AUTHOR Watson, A.
TITLE Measuring Educational Achievement in TAFE. Diploma of Teaching (Technical). Directed Study Booklet.
INSTITUTION Sydney Coll. of Advanced Education, Haymarket (Australia). Inst. of Technical and Adult Teacher Education.
PUB DATE [84]
NOTE 72p.
PUB TYPE Guides - Classroom Use - Materials (For Learner) (051)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; *Adult Education; Behavioral Objectives; Continuing Education; Foreign Countries; Higher Education; Material Development; Postsecondary Education; *Preservice Teacher Education; Program Development; *Student Evaluation; Teacher Education; Technical Education; Test Construction; Test Interpretation; Test Reliability; Test Results; Test Selection; *Test Use; Test Validity; Units of Study; Vocational Education Teachers
IDENTIFIERS Australia

ABSTRACT

This booklet was designed to help teacher education students identify and explain certain basic principles of educational measurement and apply them to the development of a valid and reliable assessment program that measures the achievement of a set of educational objectives from a curriculum that they teach. It is intended for use in the Sydney (Australia) College of Advanced Education to prepare teachers of technical and adult education. The booklet in organized under five topics. Topic 1 presents an overview of measuring educational achievement in technical and adult further education, while Topic 2 covers two principles of achievement testing: reliability and validity. In Topic 3, future teachers are shown how to plan an assessment program. Topics 4 and 5 cover constructing tests and treatment of test results. Each topic is accompanied by questions for discussion and/or student exercises. Each topic is also accompanied by a set of references and further reading. Sample tests are included in appendixes to the report. (KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED255724

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A Watson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SYDNEY COLLEGE OF ADVANCED EDUCATION

INSTITUTE OF TECHNICAL AND ADULT TEACHER EDUCATION

DIPLOMA OF TEACHING (TECHNICAL)

DIRECTED STUDY BOOKLET

"MEASURING EDUCATIONAL ACHIEVEMENT IN TAFE"

Prepared by: A WATSON

ACKNOWLEDGEMENT

This booklet has been prepared with the assistance of G Bennett and T Hay. Ideas and material have been contributed also by other colleagues and students of the TTE School in addition to the references cited.

PREAMBLE

Assessment procedures in Technical and Further Education have often come under criticism. It is not unusual to hear of students complaining about being failed in exams for not knowing work they have not been taught or about not being examined on work they have been extensively taught. It is not unusual either to hear of students hotly disputing grades assigned in essay exams, assignments and practical work.

Such criticisms reflect on the validity and reliability of TAFE assessment procedures. They come about, in part, from the fact that many teachers called upon to act as examiners have an incomplete understanding of assessment principles and techniques. Responding to these criticisms, the First Report on the Needs in TAFE in Australia (Kangan Report) recommended increased action to facilitate the "development of improved and increasingly more reliable and valid methods of assessment and reporting of students' educational achievements." (1974, p XL)

The purpose of this booklet on Achievement Testing and the related course in Educational Measurement is to assist teachers to develop the knowledge and skills necessary for these improvements to take place.

BEST COPY AVAILABLE

GENERAL OBJECTIVE

The general objective of the booklet is to have students identify and explain certain basic principles of educational measurement and apply these to the development of a valid and reliable assessment program which measures the achievement of a set of educational objectives from a curriculum which they teach.

SPECIFIC OBJECTIVES

On completion of this study booklet and the related instruction in educational measurement, students will be able to:

1. Describe the general nature of educational achievement and achievement tests.
2. Explain the role of achievement testing in the teaching/learning process.
3. Identify and describe the major types of achievement test used in Technical and Further Education.
4. Define and explain the concepts of reliability, validity and content validity as they apply to achievement testing.
5. Outline the procedures to be followed in constructing achievement tests to improve reliability and content validity.
6. Construct a table of specifications which summarises the subject matter areas and related learning outcomes for a selected course, subject, curriculum unit or module.
7. Determine appropriate forms of assessment to measure the learning outcomes and subject matter from the above table and justify these in terms of content validity and reliability.
8. Outline a comprehensive assessment program for the selected course or unit and justify the allocation of marks in relation to the table of specifications.
9. Construct multiple-choice, essay, short answer, practical and performance tests to measure the achievement of specified objectives.
10. Convert achievement test scores to grades.

CONTENT

The content of the booklet is set out as follows:

Topic 1 MEASURING EDUCATIONAL ACHIEVEMENT IN TAFE: AN OVERVIEW.

Topic 2 PRINCIPLES OF ACHIEVEMENT TESTING: RELIABILITY AND VALIDITY.

Topic 3 PLANNING AN ASSESSMENT PROGRAM.

Topic 4 CONSTRUCTING TESTS AND TEST ITEMS.

Topic 5 TREATMENT OF TEST RESULTS.

Each topic is accompanied by questions for discussion and/or student exercises. Each topic is accompanied also by a set of references and further reading.

METHOD

The booklet should be studied in conjunction with conventional classroom instruction and assignments on the topics together with a consideration of the references cited. When questions for discussion or exercises are encountered, written answers should be prepared as a basis for discussion and exercises completed.

The purpose of the questions for discussion and the exercises is to assist students to achieve the objectives of the booklet.

MAIN REFERENCE

Gronlund, N E, Constructing Achievement Tests, Prentice-Hall, New Jersey, 1977.

TOPIC 1

MEASURING EDUCATIONAL ACHIEVEMENT IN TAFE: AN OVERVIEW

OBJECTIVES

On completion of this topic, students will be able to:

1. Describe the general nature of educational achievement and achievement tests.
2. Explain the role of achievement testing in the teaching/learning process.
3. Identify and describe the major types of achievement test used in Technical and Further Education.

The topic is divided into three parts. Each part corresponds with one of the above objectives.

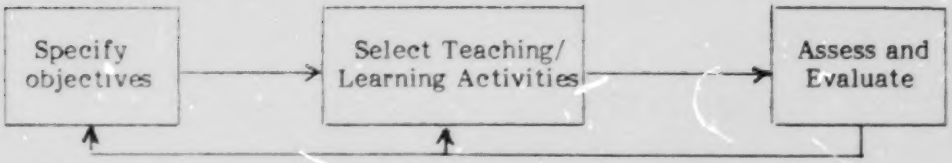
I. THE GENERAL NATURE OF EDUCATIONAL ACHIEVEMENT AND ACHIEVEMENT TESTS

Education can be thought of as a process which is basically concerned with the achievement of educational goals or outcomes. These educational goals or outcomes generally involve changes in student behaviour, and are therefore measurable. In education, as in most spheres of purposeful human endeavour, it is desirable to measure, at regular intervals, the progress that is being made towards the achievement of these goals or changes in behaviour. This is the primary function of the achievement test.

Ebel expresses this view in the following way:

> Formal education is a complex process, requiring a great deal of time and money and the co-operative efforts of many people. Efforts must be directed toward the attainment of specific goals. Education is not automatically or uniformly successful. Some methods are more effective than others. ... All of those concerned with the process of education - students, teachers, parents and school officials - need to know periodically how successful their efforts have been, so that they can decide which practices to continue and which to change. It is the function of educational measurement to provide them with this knowledge. ... The major function of a classroom test is to measure student achievement and thus to contribute to the evaluation of his educational progress and attainments. (1972, p 41)

This systematic or mechanistic view of the instructional process and the related role of assessment and evaluation may be presented diagrammatically as follows:



(Adapted from Popham & Baker, 1970, p 19)

In these terms, the chief function of the achievement test in the instructional process is to measure the student's achievement of specified educational objectives or outcomes and in this way contribute to the evaluation of his general educational progress and attainments.

QUESTIONS FOR DISCUSSION

1. Is the concept of educational achievement and the related role of achievement testing outlined above too limited?

2. What other functions do you see for assessment and evaluation in the instructional process?

II. THE ROLE OF ACHIEVEMENT TESTING IN THE TEACHING/LEARNING PROCESS

Although the measurement or summative function may be thought of as the primary purpose of achievement testing, it should be stressed at the outset that achievement tests may serve a teaching as well as a testing function and in this way have a positive or formative influence on the instructional process. As Gronlund points out, "It is easy to underestimate the role that testing plays in the teaching-learning process. ... Like teaching itself, the main purpose of testing is to improve learning, and within this larger context there are a number of specific contributions it can make." (1968, p 2)

Achievement tests may have a formative influence on the teaching-learning process in two main ways.

1. They may exert a direct influence on student learning:

   e.g.
   - aid student motivation;
   - assist retention and transfer;
   - increase self awareness and understanding;
   - provide opportunity for practice;
   - provide feedback concerning progress.

2. They provide teachers with important evaluative data necessary for effective instructional decision making.

   e.g. Data on:
   - students' progress;
   - course content and objectives;
   - teaching methods employed;
   - testing methods employed.

QUESTIONS FOR DISCUSSION

1. Comment on the formative role of achievement testing outlined above and describe any additional ways you can think of in which achievement tests may have a beneficial influence on teaching and learning.

2. It has been said that testing places students under harmful stress, exposing them to failure, destroying self confidence and joy in learning etc.

   a. Do you agree with this statement or do you feel that it is exaggerated?

   b. Under what circumstances are tests most likely to have harmful effects on learning?

III. TYPES OF ACHIEVEMENT TEST USED IN TAFE

Achievement tests used in TAFE are of various kinds and may be classified in a number of ways. This booklet will consider tests classified in terms of: A. FORM; B. METHOD OF CONSTRUCTION; and C. TYPE OF ACHIEVEMENT BEING TESTED.

A. ACHIEVEMENT TESTS CLASSIFIED IN TERMS OF FORM

The forms of achievement test commonly used in TAFE, can in turn be classified into three main divisions. These are:

1. Paper and Pencil Tests.
2. Practical and Performance Tests.
3. Assignments, Projects and Reports.

1. Paper and Pencil Tests

This is the term commonly given to the conventional test or 'exam' in which candidates write answers to questions in examination papers or examination booklets. They are often referred to as 'Theory Tests' and are typically used to measure the achievement of outcomes within the cognitive domain of educational objectives.

Gronlund further classifies paper and pencil tests as follows:

A. Supply type (Student supplies answer)

   1. Essay - extended response.
   2. Essay - restricted response.
   3. Short answer (word or phrase).
   4. Completion (fill in blanks).

B. Selection type (Student selects answer)

   1. True-false.
   2. Matching.
   3. Multiple-choice.

(Gronlund, 1968, p 7)

The chief distinguishing feature of the supply type item is that students must work out and supply the answers for themselves which are then SUBJECTIVELY scored. The selection type item on the other hand provides alternative answers for students to select from and these may therefore be OBJECTIVELY scored.

Selection type and short answer tests are generally recommended when the purpose is to test a broad range of knowledge, comprehension and perhaps application outcomes. Essay tests are generally recommended when more freedom of response is required and when the purpose is to test more complex outcomes such as the ability to analyse, integrate, organise, synthesise, evaluate etc.

(The construction of Selection and Supply type tests is described in Topic 4. Examples of each type are provided in Appendix 1 of this booklet.)

2. Practical and Performance Tests

These are the tests commonly carried out in realistic situations such as practical workshops and laboratories in which students are asked to actually perform in some way using actual tools or equipment on actual jobs, operations, experiments etc. They are typically used to measure outcomes within the Psychomotor or performance domain of educational objectives (although cognitive outcomes may also be involved).

Jones and Whittaker use the term: 'Hands-on test' to describe this type of test and differentiate it from the 'Hands-off' or paper and pencil test by saying:

> A hands-on test is one which involves the use of the tools or other hardware of the actual job. A hands-off test does not do this. Thus for a bench fitter a hands-on test may require him to produce two mating parts while a hands off test might require him to state the reading given on a diagram of a micrometer.
>
> (Jones & Whittaker, 1975, p 35)

Jones and Whittaker further classify practical or 'hands-on' tests as follows:

a. The Total Job Test in which the candidate is required to carry out the total job under real conditions - suitable for short circle and repetitive tasks and more appropriate for industrial than college training.

b. The Work Sample Test which involves the performance of jobs which are part of a total job or occupation e.g. 'removal and replacement of a clutch plate' or 'the adjustment of a tractor wheel track'.

c. The Skill Sample Test which measures the skills and operations required by a job or occupation and does not necessarily take the job situation into account. The well known engineering test piece or exercise is the classical example of this such as screw cutting, taper turning etc.

(1975, pp 36-41)

BEST COPY AVAILABLE

(The construction and set out of practical tests is described more fully in Topic 4. Examples are provided in Jones and Whittaker, 1975, pp 37, 39, 40, 42, 43 and in Appendices 2 and 3 of this booklet.)

3. Assignments, Projects, Reports etc.

Assignments, projects, reports etc may also be used to measure the achievement of course objectives. They can be distinguished from conventional tests in that they are commonly carried out in non-test situations such as home studies, libraries, resource centres, laboratories etc. under non-test conditions.

They are legitimate forms of assessment however and are often used to measure attainment of the more complex cognitive outcomes such as analysis, synthesis, and evaluation or a combination of cognitive, psychomotor and affective outcomes. They are often preferred because they bring reality and therefore validity to the test situation and they are suitable for progressive and cumulative testing.

QUESTIONS FOR DISCUSSION

1. Is the above classification of achievement tests in terms of form comprehensive enough to account for most of the types of test commonly used in TAFE or have certain important types of test and learning outcomes been overlooked?

2. Look at the examples of the major types of paper and pencil test provided in Appendix 1. What do you see as the chief uses, advantages and disadvantages of the major types of paper and pencil test?

B. ACHIEVEMENT TESTS CLASSIFIED ACCORDING TO THE METHOD OF CONSTRUCTION

Achievement tests may also be classified in terms of the method of construction. Two classifications are particularly relevant for TAFE.

1. Standardised Vs Teacher Made Tests

A standardised test is one that:

a. has been carefully constructed by experts with tryout and revision;

b. has explicit instructions for administration;

c. has tables of norms or standards for the interpretation of scores.

(Ebel, 1979, pp 310-311)

A teacher made test is one constructed by a classroom teacher or group of teachers and it rarely has the above features.

2. External Vs Internal Tests

External or centrally set tests are those constructed by a central authority such as a departmental school or division or an external regulatory body. They provide common exam content, uniform administration and marking procedures and scoring standards across a wide range of classes and colleges. They could be standardised or teacher made tests and they are often used for final or end of year examinations.

Internal tests are those set by individual teachers for particular classes and they may differ in content, procedure, marking and scoring standards from class to class. They are often used to ascertain so called 'classmarks' and are often used also for progressive or cumulative testing through the year.

External tests have been criticised for their external control over the curriculum and their tendency to encourage teachers to 'teach to the test'. Internal tests have also been criticised for their lack of uniformity in terms of content, marking, scoring standards etc.

(A further discussion of the issues and examining procedures involved in relation to centrally set vs internal tests and end of year vs progressive tests can be found in Hermann, 1969).

QUESTIONS FOR DISCUSSION

1. How would you defend both types of test against the above criticisms?

2. What are the particular advantages of each type of test?

3. What compromise system would you advocate so as to make use of the advantages of each type of test and minimise the disadvantages?

C. ACHIEVEMENT TESTS CLASSIFIED IN TERMS OF THE TYPE OF ACHIEVEMENT BEING TESTED

The types of achievement test so far described may be used to measure three different types of student achievement. That is, for different purposes or reasons. These tests may be classified as follows:

1. The Diagnostic Test

Diagnostic tests are used to measure: (a) previous achievement or readiness for a course; or (b) underlying causes of lack of achievement.

The purpose of diagnostic testing is to (a) place the student properly at the outset of instruction, or (b) discover the underlying causes of lack of achievement so that remedies may be applied as instruction unfolds. (Bloom et al, 1971, p 87).

The former type are sometimes referred to as placement or entry and pre tests. Diagnostic tests are commonly standardised tests, although they may be teacher made. They generally contain a large number of items on specific areas and each test is generally built around common errors made by students rather than on a broad sample of course objectives. (Gronlund, 1968, p 9).

BEST COPY AVAILABLE

NOTE: The construction of diagnostic tests is beyond the scope of this booklet.

2. The General Achievement Test

General achievement tests are used to measure general or relative achievement over a broad range or sample of course objectives and to rank or compare students in terms of their achievement of these objectives.

The chief purpose of the general achievement test is to grade or rank the learner. On the basis of these tests students are passed or failed, selected or dropped, given an A, B, C, D or E grade and results are reported to parents and administrators. General achievement tests are directed at a "more general assessment of the degree to which the larger outcomes have been attained over the entire course or some substantial part of it." (Bloom et al, 1971, p 61).

> The general achievement, or survey test, (writes Gronlund) is designed to measure a broad range of learning outcomes and to rank pupils in the order of their achievement. Comprehensiveness and representativeness of sampling are important features, as is the difficulty of the test. To obtain a dependable ranking of students, in terms of the degree to which they are achieving the instructional objectives, the test should adequately sample all aspects of the course and be difficult enough to spread students out in accordance with their relative attainment of the learning outcomes. (1968, p 9).

This type of testing is often referred to as norm-referenced testing because grades tend to depend on averages or standards within the group tested. It tends to stress the summative or judgemental function of testing.

Most centrally set and end of year or final exams in TAFE are general achievement tests. They may be standardised or teacher made and they may contain pencil and paper tests, practical or performance tests and assignments, projects and reports.

Many of the references on test construction tend to concentrate on this form of test.

3. The Mastery Test

Mastery tests are used to measure absolute achievement or mastery of a finite and clearly specified set of objectives and to ascertain which students have or have not mastered these objectives.

The purpose of mastery testing is to:

> ... determine the degree of mastery of a given learning task and to pinpoint the part of the task not mastered. ... The purpose is not to grade or certify the learner; it is to help both the learner and the teacher focus upon the particular learning necessary for movement toward mastery. (Bloom et al, 1971, p 61).

This type of testing is often referred to as criterion-referenced testing as student performance is measured against a set standard or criterion and grades tend to be pass or fail, reflecting achievement or non-achievement of the criterion.

Data from this form of test tends to be used for the more formative or positive functions of testing and instructional decision making. As Gronlund puts it, the purpose of formative testing is to "monitor learning progress." (1977, p 19).

Mastery tests are commonly used in TAFE when there is a minimum or basic set of essentials to be mastered by all students, e.g. on the completion of a module or unit of instruction or on the completion of a practical job or operation. Tests have to be constructed so that all objectives are tested. The individual items or parts of the test may have a low level of difficulty although there is often a high pass mark for the total test.

Mastery tests are commonly internal and teacher made and are often used in an informal way in the form of revision quizzes etc. They may also be standardised and centrally set and incorporated in quite a formal way into a modular curriculum for example. They may include paper and pencil, practical and assignment elements.

NOTE: The principles of test construction outlined in the remainder of this booklet may be applied to the construction of a Mastery and General Achievement test but not to Diagnostic tests.

QUESTIONS FOR DISCUSSION

1. It has been suggested that the traditional general achievement or norm-referenced test used to test broad areas of achievement encourages a negative approach to testing with emphasis on selection and placement rather than guidance and development and that these should be replaced by the regular use of mastery or criterion-referenced testing which indicates "not whether Johnny knows more than Mary, but exactly which objectives of instruction have been achieved and which have not." (Ebel, 1979, pp 5-6).

   a. Do you agree with these criticisms of the general achievement test?

   b. Will this form of test need to be retained? If so, for what reasons?

   c. What difficulties do you see in a system which breaks down the curriculum into small units and uses mastery tests to test the achievement of each unit?

2. Why is it necessary for TAFE teachers to be familiar with the various types of achievement test which may be used and to understand their functions?

REFERENCES AND FURTHER READING

BLOOM, B S, HASTINGS, J T and MADAUS, G F. Handbook on Formative and Summative Evaluation of Student Learning, McGraw-Hill, New York, 1971.

BROWN, F G, Principles of Educational and Psychological Testing, Dryden Press, 1970, Ch 10; Holt, Rinehart & Winston, 1976, Chs 13, 14.

EBEL, R L, Essentials of Educational Measurement, Prentice-Hall, New Jersey, 1972, Ch 2; 1979, Chs 1 and 2.

GRONLUND, N E, Constructing Achievement Tests, Prentice-Hall, New Jersey, 1st Edn, 1968, Ch 1; 2nd Edn, 1977, Ch 1.

HERMANN, G, 'Examining Systems in Technical Education', The Australian Technical Teacher, May, 1969, pp 29-35.

JONES, A and WHITTAKER, P, Testing Industrial Skills, Gower Press, UK, 1975, Ch 4.

POPHAM, W J and BAKER, E L, Establishing Instructional Goals, Prentice-Hall, New Jersey, 1970.

THEOBOLD, J. Classroom Testing: Principles and Practice, Longman, Aust, 1974, Ch 1, pp 1-2.

TOPIC 2

PRINCIPLES OF ACHIEVEMENT TESTING: RELIABILITY AND VALIDITY

INTRODUCTION

The previous topic outlined a range of achievement tests commonly used in TAFE and the purposes for which they are used. The three broad functions stressed were:

1. The measurement of student achievment.
2. The formative influence on the teaching/learning process.
3. The contribution of evaluative data for instructional decision making.

If achievement tests are to fulfil these functions properly and effectively however, they need to be constructed in such a way that:

a. student grades and evaluative data are reliable and dependable - and not based on distorted or biased information; and

b. testing, teaching and subsequent learning are directed towards the important and worthwhile aspects of a course rather than towards the trivial or limited aspects.

If these standards are to be achieved, two important principles or criteria of achievement testing must be followed. These are referred to as:

1. RELIABILITY and
2. VALIDITY

The purpose of this topic is to explain these principles as they apply to achievement testing and outline the procedural steps in test construction which must be followed if the principles are to be achieved.

SPECIFIC OBJECTIVES

On completion of this topic, students will be able to:

1. Define and explain the concepts of reliability, validity and content validity as they apply to achievement testing.
2. Outline the procedures to be followed in constructing achievement tests to improve reliability and content validity.

I. RELIABILITY

a. What is Test Reliability?

Reliability can be thought of as the consistency or precision of a test - the extent to which test scores measure whatever they do measure consistently and precisely and are free from error or bias.

Ebel defines test reliability as "the consistency with which a set of test scores measures whatever they do measure." (1972, p 409)

An example of a reliable measuring instrument would be a steel rule, because it yields measurements which are:

1. relatively consistent from measurement to measurement and occasion to occasion;
2. relatively precise or free from error.

(a micrometer would be a more reliable instrument).

A reliable test, like a steel rule, should yield results which are:

1. relatively consistent irrespective of the time of test or the marker;
2. relatively precise or free from error.

b. Methods of Estimating Test Reliability

There are various statistical methods of estimating the reliability of a test. They generally involve test-retest procedures to estimate consistency over time with the same test or between equivalent forms of a test. There is also a method of estimating internal consistency within the one test. (These methods are described adequately in Gronlund, 1977, p 139).

Using these methods, factors and procedures have been identified which can be shown to influence the degree of test reliability for a given test.

c. Factors Which Impair Test Reliability

Test reliability may be impaired and scores distorted by the following factors.

1. Subjective or unreliable marking - marking which is subject to the bias or condition of the marker. The effects of this factor are most severe on essay tests, practical or peformance test, and project and assignments.
2. Poorly worded and ambiguous questions - questions which leave open a number of interpretations, cloud the real intent of the question or are too easy or difficult. The effects of this factor can be severe on both essay (supply) and objective (selection) type questions.

3. Chance factors such as lucky guesses, lucky choice of topics or questions, condition of the student at the time of the test etc. These may distort scores on all types of test.

d. Procedures to Improve Test Reliability

Certain procedures may be followed in test construction and in the treatment of test results that will generally improve test reliability. These are as follows:

1. Improve Marker Reliability. This can be achieved through the use of good marking guides for essay questions, practical tests and assignments or by the use of objectively scored questions such as multiple-choice and true-false questions.

2. Word questions carefully so as to:

   - remove ambiguities;
   - clarify the response(s) required;
   - eliminate questions which are too easy or too difficult.

3. Reduce the impact of chance factors by such means as:

   - longer more comprehensive tests;
   - cumulative or progressive testing.

In relation to the wording of questions, it has been found that regular analysis of student responses on objectively scored tests, using a procedure called Item Analysis, is very useful in detecting questions which:

a. are too easy;
b. are too difficult;
c. fail to discriminate consistently between high and low achievers.

(A procedure for carrying out the item analysis procedure is described in Gronlund, 1977, pp110-116.)

In this regard, it has been shown statistically that the reliability coefficient of test scores will be greater for a test:

a. composed of more homogeneous items than from a more heterogeneous test;

b. composed of more discriminating items than from a test composed of less discriminating items;

c. whose items are of middle difficulty than from a test composed mainly of quite difficult or quite easy items.

(Ebel, 1972, p 427)

BEST COPY AVAILABLE

II. VALIDITY

a. What is Validity?

Validity is often explained in relation to reliability.

Whereas reliability can be thought of as the extent to which a test measures consistently and with precision whatever it is measuring; validity should be thought of as the extent to which a test measures precisely and accurately what it sets out to measure.

Whereas reliability is concerned with the question: 'How well is the test measuring?' Validity is more concerned with the question: 'What is the test measuring?'

In this regard, Ebel defines validity as "the accuracy with which a set of test scores measure what they ought to measure." (1972, p 409)

Ebel then goes on to differentiate validity from reliability in the following way:

> If the perforations on a target made by successive shots from a rifle are all clustered closely, the rifle is performing reliably. If those perforations are all clustered in the bull's eye, the rifle is also performing validly.
>
> The number of words in a poem can be measured with high reliability. Such a measure would be readily accepted as a valid measure of the length of the poem. It would not, however, be accepted by most poets or literary critics as a valid measure of the literary merit of the poem.
>
> (1972, p 409)

b. Characteristics of Validity

The concept of validity as used in testing, can be clarified further by noting the following general points or characteristics.

1. Validity refers to the interpretation of test results (not the test itself).
2. Validity is inferred from available evidence (not measured).
3. Validity is specific to a particular use (selection, placement, evaluation of learning etc.)
4. Validity is expressed by degree (e.g. high, moderate, low).

(Gronlund, 1977, p 131)

c. Basic Types of Validity

A standard and widely adopted classification system (American Psychological Association, 1974) divides validity into four basic types: content validity; predictive validity; concurrent validity; construct validity. (These are described in Gronlund, 1977, p 131 and in most texts on this topic.)

The test construction procedures discussed in the remainder of this booklet relate specifically to content validity.

III CONTENT VALIDITY

a. What is Content Validity?

Content validity can be thought of as the extent to which a test measures a representative sample of course content.

A test which has a high degree of content validity would measure the major or important outcomes of a course rather than the unimportant or trivial ones and would measure them in a way which accurately reflects their importance or emphasis in the course curriculum.

Content validity is concerned with the question: 'How adequately does the test content sample the total subject matter to be assessed?'

A test which is high on content validity would be made up of items which form a balanced sample of the total subject matter and learning outcomes of the course or subject being assessed.

Gronlund suggests as an example a spelling test containing a sample of 50 words from a 500 word spelling list. He maintains that such a test would have high content validity "to the degree to which it provided an adequate sample of the 500 words it represented." (1977, p 132)

b. Factors Which Influence Content Validity

Although there is no simple statistical method to determine content validity, it can be demonstrated that content validity will tend to be high for tests which observe the following basic principles:

1. Measure clearly defined learning outcomes which are in harmony with instructional objectives.

2. Measure a representative sample of the learning outcomes and subject matter of the course or subject.

3. Include the forms of assessment which are most appropriate for measuring the desired learning outcomes.

4. Have marks allocated in such a way as to accurately reflect the structure and emphases in the course being tested and the way it was taught.

(adapted from Gronlund, 1977, pp 7-13)

c. Procedures to Improve Content Validity

Accordingly a number of basic procedures in test construction can be identified which will improve content validity.

These are:

1. Identify the subject matter topics and related learning outcomes to be measured.

2. Construct a table of specifications which summarises these topics and learning outcomes and the relative weightings of each.

3. Select types of test and test item which are most appropriate for measuring the desired learning outcomes.

4. Plan the total test or assessment program so that mark allocations accurately reflect the importance of topics and outcomes being tested.

5. Write the tests and test items so that they do in fact measure the outcomes which they are intended to measure.

(adapted from Gronlund, 1977, p 132)

The remainder of this booklet will describe in more detail the principles and procedures which should be followed in the construction of a valid and reliable achievement test or assessment program.

Generally, the first procedures followed are those to ensure content validity. Subsequently measures are taken to improve test reliability.

QUESTIONS FOR DISCUSSION

1. Explain test reliability, validity, and content validity in other words. (Use definitions and/or explanations and examples not given here).

2. Rank the following measuring instruments in terms of their reliability, and justify your ranking.

   a. steel rule;
   b. micrometer;
   c. essay test;
   d. multiple-choice test.

3. Give an example of a test or test item with poor content validity. Explain the likely consequences of this for students' results.

4. What other qualities of a test are important in addition to reliability and validity?

REFERENCES AND FURTHER READING

BROWN, F G, Principles of Educational and Psychological Testing, Dryden Press, 1970, Chs 3, 4, 5, 6; Holt, Rinehart & Winston, 1976, Chs 4 and 5.

EBEL, R L, Essentials of Educational Measurement, Prentice-Hall, New Jersey, 1972, Chs 15 and 16; 1979, Chs 14 and 15.

GRONLUND, N E, Constructing Achievement Tests, Prentice-Hall, New Jersey, 1968, Ch 8; 1977, Ch 9.

THEOBOLD, J. Classroom Testing: Principles and Practice, Longman, Aust, 1974, Ch 5, pp 53-57.

TOPIC 3

PLANNING AN ASSESSMENT PROGRAM

INTRODUCTION

The previous topic explained the concepts of reliability and validity as they apply to achievement testing and outlined a number of principles and procedures to be followed in test construction to improve test reliability and content validity.

The purpose of the present topic is to describe these procedures (and related principles) more fully and apply them to the planning of a reliable and valid assessment program.

The planning of an assessment program can be accomplished of course in a number of ways. Most professional test constructors and classroom teachers would agree, however, that the following four procedures are beneficial if not essential:

I Determine the purpose of the tests to be used.

II Identify the subject matter and learning outcomes to be assessed on a table of specifications.

III Determine the most appropriate forms of assessment and assessment strategies for the content being assessed.

IV Outline the overall program so that the forms of assessment used and mark allocations faithfully reflect the course or unit being assessed.

These procedures are directed primarily at ensuring content validity. Measures to improve test reliability are taken as the planning and item writing proceed.

The content of this topic will focus on these four procedures in turn.

SPECIFIC OBJECTIVES

On completion of this topic students will be able to:

1. Construct a table of specifications which summarises the subject matter areas and related learning outcomes for a selected course, subject, curriculum unit or module.

2. Determine appropriate forms of assessment to measure the learning outcomes and subject matter from the above table and justify these in terms of content validity and reliability.

3. Outline a comprehensive assessment program for the selected course or unit and justify the allocation of marks in relation to the table of specifications.

1. DETERMINING THE PURPOSE OF THE TESTS TO BE USED

It was noted in topic 1, that achievement tests used in TAFE take many forms and are used for different purposes in the instructional program. It is important to realise that different test characteristics may necessitate some modifications in test design.

By way of reminder, the two types of achievement test of major concern in this booklet, together with their characteristics important for test construction, are summarised in the following table.

Table 3.01 Characteristics of Two Types of Test, Classified in Terms of Function

Type of Test	Function of Test	Sampling Considerations	Item Characteristics
Mastery or Formative	Monitor learning progress or absolute achievement of a narrow range of unit objectives.	Include all unit objectives if possible or those most essential.	Items match difficulty of unit objectives and are criterion-referenced.
General Achievement or Summative	Measure relative achievement of a broad range of course objectives.	Select representative sample of course objectives.	Items are generally of median difficulty and norm-referenced.

Adapted from Gronlund, 1977, p 19.

Before the actual planning begins, therefore, the test constructor must decide whether he intends to use mastery tests, general achievement tests or both in the assessment program.

The procedures to be described in this topic will be applied to both mastery and general achievement tests. They may also be applied to standardised, centrally set tests or internal, teacher made tests.

The procedures will be applied to the planning of a comprehensive assessment program made up of a number of forms of assessment. They may be applied also to the planning of a single achievement test. Decisions about the precise forms of assessment to be used should not be made, however, until the content and related learning outcomes to be assessed are identified.

BEST COPY AVAILABLE

QUESTIONS FOR DISCUSSION

1. What are the chief implications for the test planner when constructing an assessment program which has as its main purpose the use of mastery tests?

2. What are the implications when the test planner intends to use general achievement tests.

II. IDENTIFYING THE CONTENT TO BE ASSESSED: THE TABLE OF SPECIFICATIONS

The first major procedure to be followed in the planning of an assessment program, once the purpose of the program has been determined, is to identify and specify what is to be measured; that is, the subject matter and related learning outcomes, to be assessed.

The principle being followed here is that a test will have high content validity to the extent that it measures a truly representative sample of subject matter and learning outcomes from the course or curriculum unit being assessed.

Such a sample can be achieved by following three related steps:

Step 1. Outline the subject matter areas to be assessed.

Step 2. Identify the related learning outcomes to be assessed.

Step 3. Construct a table of specifications which summarises:

a. the subject matter areas to be assessed;
b. the learning outcomes to be assessed;
c. the relative weighting (or proportion of marks) to be given to each subject matter area and learning outcome in the total assessment program.

Step 1. Outline the Subject Matter to be Assessed

The first step in obtaining a representative sample of course content is to outline the subject matter divisions or topics of the course or unit to be assessed.

Whereas the subject matter of a course may be listed in detail for curriculum or teaching purposes, only the major categories or headings need to be listed as a basis for test planning.

The amount of detail included of course will depend on the nature and scope of the assessment program. For a general achievement test of a whole course or subject, only the major subject matter headings should be included to prevent the outline becoming unmanageable. For a mastery test of a discrete unit or module, more detail on topics and sub-categories may be included.

In addition, each subject matter division should be given a weighting. These weightings, usually expressed as percentages, indicate the relative proportion of the total weighting or mark allocation to be given to each division. Weightings will reflect such factors as time spent in instruction on each division, the relative importance of each division etc.

For examples of subject matter outlines see tables 3.02 and 3.03.

Table 3.02 Outline of Subject Matter and Related Weightings for a Module in Basic Electrical Fitting

Topics	Weightings (%)
Safety Procedures	5
Hand Tools	30
Measuring/Marking Tools	30
Machine Tools	30
Stores Procedures	5
	100

Table 3.03 Outline of Subject Matter and Related Weightings for a Unit Course in Biology

Topics	Weightings (%)
The Cell	30
Nutrition	25
Respiration	15
Reproduction	18
Ecology	12
	100

It should be re-emphasised that these subject matter outlines will vary depending on the scope of the curriculum unit being assessed.

An outline for the total first year course in Applied Electricity for example, might include Basic Electrical Fitting as only one of its topics. Conversely, an outline for a single unit in Biology on The Cell would include the components and aspects of The Cell treated.

Sometimes the details or sub-categories for a division may be increased to indicate its importance.

The weightings allocated to each division will be influenced, of course, by the biases of the test planner.

Nevertheless, as far as possible, the outline should be a faithful representation or sample of the subject matter of the course or unit as it was taught.

Step 2. Identify the Learning Outcomes to be Assessed.

Once the subject matter divisions have been outlined, the next step is to identify the learning outcomes which relate to each division. These learning outcomes should accurately reflect the objectives of instruction as set out in the course or curriculum unit being assessed and they should be stated in a manner useful for testing.

If these outcomes are not identified and stated clearly in the curriculum, a useful starting point is the Taxonomy of Educational Objectives (Bloom et al, 1956; Krathwohl et al, 1964; Harrow, 1972.) According to this scheme, educational outcomes are of three main types and can be classified into three broad domains:

1. The Cognitive Domain, concerned with knowledge outcomes and intellectual abilities and skills.

2. The Psychomotor or Performance Domain, concerned with motor and general skills (practical, laboratory, communication, work study, managerial skills etc.)

3. The Affective Domain, concerned with attitudes, interests, appreciations etc.

Although there will be variations from course to course, most sets of instructional objectives can be classified in terms of these three broad domains.

In addition, each of these broad domains is subdivided into sub-categories of general learning outcomes of different levels of complexity and each outcome is described in terms of the specific student behaviours that are to be accepted as evidence that the outcome has been achieved.

(The categories in each domain are summarised in Appendix 5 and some action verbs or specific behaviours associated with each domain are provided in Appendix 6.)

It should be emphasized, however, that in setting out the learning outcomes to be assessed, the test planner should employ the categories and terminology which best reflect the objectives of the course or unit being assessed.

For examples of sets of learning outcomes see Tables 3.04 and 3.05.

Table 3.04 Learning Outcomes to be Assessed in Basic Electrical Fitting

COGNITIVE DOMAIN	
KNOWLEDGE:	Identify, name, describe tools and their parts, list uses.
COMPREHENSION:	Explain functions, advantages etc.
APPLICATION:	Describe the use, care and operation.
AFFECTIVE DOMAIN	
ATTITUDE DEVELOPMENT:	Demonstration of safe and conscientious work habits.
PERFORMANCE DOMAIN	
PRACTICAL SKILLS:	Measure, mark out, drill, file, cut, etc.

Table 3.05 Learning Outcomes to be Assessed in Biology

COGNITIVE DOMAIN	
KNOWLEDGE:	Identify, name, define, describe.
COMPREHENSION:	Classify, explain, predict, distinguish between.
APPLICATION:	Compute, solve, relate.
ANALYSIS	Differentiate, diagram, subdivide.
EVALUATION:	Compare, justify, conclude.
PERFORMANCE DOMAIN	
LABORATORY SKILLS:	Perform basic laboratory operations and procedures.

It should be noted that while the taxonomies provide a useful guide or check list for identifying outcomes and a useful vocabulary for describing and classifying them, only the test planner can determine which outcomes are to be assessed and which are not after close scrutiny of the course objectives. The outcomes included in a test program will be determined by such factors as the specific nature of the course, the philosophy of the school, the types and needs of students as well as the purpose of the test program. For a general achievement test a wide sample of objectives would be included whereas a mastery test might confine itself to quite a narrow band of outcomes.

The taxonomies are neutral also in relation to the relative emphasis or weighting to be given to each outcome. This again will be influenced somewhat by the biases of the test planner, but as far as possible weightings allocated should reflect the relative emphasis given to the different outcomes as the course was taught.

Learning outcomes may be related to each subject matter division and weightings allocated at this point or this may be left until the construction of the table of specifications.

Step 3. Construct a Table of Specifications

When the learning outcomes have been identified and defined and subject matter divisions outlined, the test planner can begin the task of planning the test or test program. The dual consideration of subject matter and learning outcomes provide the test planner with a convenient format for constructing a test plan. This format is referred to as a table of specifications.

The table of specifications is a table which relates outcomes to be tested to subject matter and indicates the relative weightings to be given to each. Its purpose, as was noted earlier, "is to provide assurance that the test will measure a representative sample of the learning outcomes and the subject-matter topics to be measured". (Gronlund, 1977, p 27).

Tables of specifications may be constructed for single tests or comprehensive assessment programs, for courses, subjects, curriculum units or modules. They will vary in detail and specificity according to the scope of the course or unit being assessed and the purpose of the test or assessment program.

Four examples are provided below and three simple steps of construction will be described. (For a more comprehensive account and additional examples the reader should consult Bloom, Hastings and Madaus, 1971, Ch 23).

Table 3.06 Table of Specifications for a Module in Basic Electrical Fitting

Outcomes / Topics	Cognitive Domain			Performance Domain	Affective Domain	Total
	Knowledge (name identify)	Comprehension (Explain)	Application (Describe)	Practical Skills	Attitude Development	Emphasis
Safety Procedures	3	2	-	-	-	5
Hand Tools	4	3	3	15	5	30
Measuring/ Marking Tools	4	3	3	15	5	30
Machine Tools	4	3	3	15	5	30
Stores Procedures	3	-	2	-	-	5
Total Emphasis	18	11	11	45	15	100
		40			60	

Table 3.07 Table of Specifications for a Unit Course in Biology

Outcomes / Topics	Cognitive Domain			Performance Domain	
	Knowledge	Comprehension	Application Analysis, Evaluation	Laboratory Skills	Total Emphasis
Cell	8	12	5	5	30
Nutrition	7	8	5	5	25
Respiration	5	4	3	3	15
Reproduction	6	4	4	5	18
Ecology	5	2	3	2	12
Total Emphasis	36	30	20	20	100%

(Adapted from Theobold, 1974, p 29).

BEST COPY AVAILABLE

Table 3.08 Table of Specifications for a Curriculum Unit: Basic Camera Techniques

Outcomes / Topics	Cognitive Domain				Performance Domain	Total Emphasis
	Knowledge	Comprehension	Application	Evaluation	Practical Skills	
Camera Types	4	-	-	-	-	4
Image Formation	6	3	6	-	-	15
Photographic Routine	4	1	-	-	4	9
Standardisation	4	1	-	-	4	9
Line Copy and Negative Assessment	3	1	-	1	-	4
Exposure Control	4	1	-	-	-	5
Fixed Aperture System	5	-	5	-	8	18
Fixed Ratio System	5	-	5	-	8	18
Development	2	1	-	2	4	9
Emulsion Characteristics	2	1	-	2	4	9
Total Emphasis	38	9	16	5	32	100

Table 3.09 Table of Specifications for The Subject: Individual Development (From Typing and Communication)

Outcome / Topics	Cognitive Domain			Affective Domain			Total Emphasis
	Knowledge	Compre-hension	Appli-cation	Respon-ding	Valuing	Organi-zing	
Exercise and Relaxation	6	6	12	-	-	-	24
Diet and Nutrition	6	6	12	-	-	-	24
Recreation and Leisure	6	6	10	-	-	-	22
Self-Peception	-	-	-	9	4	2	15
Inter-Personal Relationships	-	-	-	9	4	2	15
Total Emphasis	18	18	34	18	8	4	100
		70			30		100

There are three simple steps in the construction of a table of specifications.

a. Down the left hand side or vertical axis of the table list the subject matter divisions or topics to be assessed. These will generally be the divisions previously outlined. More or less detail may be added as required.

b. Along the top or horizontal axis of the table list the learning outcomes to be assessed. These may be grouped into the major objective domains or categories and it is usual to use the more general learning outcomes for these headings. More specific behaviours may be indicated for each outcome if this detail is required. (At this stage it is useful to place small ticks in intersecting cells to indicate where outcomes do and do not relate to the subject matter divisions).

c. Allocate a weighting to each subject matter and outcome category. These weightings written as percentages indicate the relative emphasis to be given to each topic and outcome and, consequently, indicate also what proportion of the total mark allocation is to be devoted to each category.

One method of approaching the allocation of weightings task is to first enter in the vertical total column (far right) the weightings (previously determined) for each subject division or topic . The total weighting for each topic can then be distributed across the rows, allocating a proportion of that weighting to each outcome which relates to that topic. The total weighting for each outcome may then be calculated and entered along the bottom row.

Remember that these weightings represent the judgement of the test planner as to the relative emphasis to be given to each outcome category. The weightings for marginal categories may have to be modified and compromises made before totals are finally settled. As near as possible, though, the weightings finally assigned should reflect the emphases given to the various outcomes of the course as it was taught.

In summary, the table of specifications indicates the topics and outcomes to be covered by the test or assessment program and the relative emphasis to be given to each topic, outcome and topic-outcome category. Once constructed, the table should serve as the basis for the planning of the total assessment program. On the basis of the table of specifications, appropriate forms of assessment may be selected, and mark allocations and the number of items required for each table category can be estimated.

To the extent that the table of specifications is a faithful representation of the course as it was taught and to the extent that the assessment program when it is constructed accurately mirrors the categories and emphases in the table, then the program will have content validity.

(Brown, 1970, p 262)

EXERCISES

1. Prepare an outline of subject matter divisions or topics from a course, subject, curriculum unit or module which you wish to assess.

2. Allocate weightings for each topic which reflect your estimation of the relative importance of each topic in the course or unit etc.

3. Outline the learning outcomes to be assessed in relation to the above subject matter. These learning outcomes should be grouped into appropriate classifications and stated in terms useful for testing.

4. Construct a table of specifications which summarises the above subject matter divisions and learning outcomes and indicates the relative emphasis to be given to each topic, outcome and topic-outcome category in the total assessment program. (If necessary you may prepare several tables showing the various stages of construction).

III. DETERMINING THE FORMS OF ASSESSMENT TO BE USED

The table of specifications summarises the learning outcomes which the test constructor wishes to assess in relation to each topic or subject matter division of the course or unit under consideration.

The assessment program can be thought of as a device or strategy to bring out those outcomes or specified behaviours so that judgements can be made concerning the standard of their achievement.

The next procedure to be followed in planning the assessment program, therefore, is to determine the forms of assessment and assessment strategies to be used. Brown makes this point quite strongly in the following way.

> The instructor's educational objectives for the course will also determine the content and format of his exams. If his objectives are to teach basic vocabulary and facts, his test can consist, say, of purely factual items plus definitions; if the main objective is to teach a particular computational skill, the exam should consist of applications of that skill . . . of course, in most circumstances, there will be several objectives and a comprehensive examination (or assessment program) will have to assess several distinct types of performance. . .
>
> (Brown, 1920, p 200)

The principle to be followed is that to ensure content validity, the forms of assessment and assessment strategies employed should be those best suited to bring out and assess the learning outcomes and behaviours specified in the table of specifications.

Again this may be achieved by following a number of simple steps.

Step 1. Determine the forms of Paper and Pencil Test to be Used

It was noted earlier that paper and pencil tests are commonly used to measure the cognitive outcomes of a course or unit. The task for the test constructor is to determine those outcomes from the table of specifications which can be best measured by paper and pencil tests and then select the types of test which give maximum validity and reliability.

In Table 3.10 the two broad classifications of paper and pencil test are compared on six criteria. Each of these criteria relates in some way to content validity and/or reliability.

Table 3.10 A Comparison of the Two Main Categories of Paper and Pencil Test

CRITERION	SELECTION TYPE - True False - Multiple Choice - Matching	SUPPLY TYPE - Extended Essay - Restricted Essay - Short Answer
1) Taxonomy Outcomes Measured	Best suited to measure knowledge, comprehension and application outcomes.	Best suited to measure higher mental processes: analysis; synthesis; evaluation. May also measure the ability to: (1) organise knowledge, and ideas. (2) express opinions, attitudes and appreciations. (3) show initiative, originality and creativity.
	Measure knowledge of facts and basic mental abilities efficiently permitting no evasion. However, may emphasise superficial learning and recognition of the correct answer rather than remembering and understanding.	But do not measure purely factual information efficiently.
2) Control Over Behaviours	If questions constructed carefully, maximum control over behaviour and knowledge tested.	Very little control over behaviour and content tested. Open to a variety of interpretations. Students can avoid the issue. More control as response is restricted.

CRITERION	SELECTION TYPE	SUPPLY TYPE
3) Scope of Content Tested	Large number of items makes possible extensive sampling of content in one test and in a relatively short time.	Cover only a limited field of knowledge in any one test. Essay questions take so long to answer that relatively few can be answered in a given time period.
4) Ease of Preparation	Require writing many questions and extreme care must be taken to avoid ambiguities or giving clues to the correct answer.	Require writing only a few questions but this may be done quickly and carelessly resulting in questions becoming ambiguous, too general and of unequal difficulty.
5) Marking	Objective, quick and highly reliable. (High marker reliability).	Time consuming, difficult, subjective and less reliable. (Low marker reliability).
6) Factors Distorting Subject's Scores	Reading ability. Guessing. Poor item construction.	Writing ability. Punctuation. Grammar. Cleverness in written expression (bluffing). Halo effects.

(Adapted from Gronlund, 1977, p 78).

Step 2. Determine the Practical and Performance Tests to be Used

Outcomes in the psychomotor or performance domain (motor or general skills) are best measured by some form of performance or 'hands-on' test. These will probably be of the work sample or skill sample variety and will be planned to occur in a workshop, laboratory or simulated 'real-life' situation.

To achieve content validity the tests will need to be constructed so as to highlight the abilities and skills in question and to improve reliability, marking scales and guides will be required.

Outcomes in the affective domain may also be measured by performance tests.

Step 3. Determine Whether Assignments, Projects and/or Reports are to be Used.

Assignments, projects and/or reports may be used to measure the higher, more complex cognitive outcomes, general performance outcomes, affective outcomes, or a combination of outcomes.

They have the advantage of reality and therefore content validity, in that they can be related very directly to course objectives. Validity and reliability may be adversely affected, however, by careless design and wording and by subjective marking.

These forms of assessment are very useful also for cumulative or progressive testing.

Step 4. Determine Whether Centrally Set, External or Teacher Made, Internal Tests are to be Used.

If the chief function of the assessment program is testing for mastery with an emphasis on the formative aspects of testing, then a progressive series of internal teacher-made mastery tests will probably be used.

If the function, however, is to measure general achievement with an emphasis on the summative aspects of testing, some form of centrally set final or general achievement test will need to be used at the completion of each unit.

In actual practice however, a compromise is usually arrived at and the assessment program is made up of some combination of the various forms of assessment and assessment strategies. That is, a combination of: internal and centrally set tests; final and progressive tests; paper and pencil tests, practical tests, assignments, projects and reports.

Step 5. Outline the Forms of Assessment to be Used to Assess the Various Learning Outcomes.

In order to assist in the subsequent allocation of marks, it is useful at this point to outline the forms of assessment and assessment strategies to be used along with the learning outcomes to be assessed by each. It may be useful also to list in point form the considerations which underlie and justify the various decisions made. This can be done along the lines of Table 3.11.

EXERCISES

1. Prepare an outline of the forms of assessment and assessment strategies you would use to measure the achievement of the learning outcomes specified in your table of specifications.

2. Justify each of the above in terms of content validity and reliability (in point form).

Table 3.11 Forms of Assessment and Assessment Strategies to be Used to Assess Learning Outcomes in Biology.

OUTCOMES TO BE ASSESSED	FORMS OF ASSESSMENT	JUSTIFICATION
Knowledge and Comprehension	Objective Tests Selection Type - Multiple Choice - True/False Comprising: - one centrally set test (end of course) - five teacher made class tests (progressive throughout year).	. Objective tests best suited to measure knowledge and understanding outcomes. . Centrally set tests required to ensure uniformity of standards. (Large number of part-time teachers, not familiar with expected standards) . Teacher-made tests for regular feedback on student progress.
Higher Mental Abilities . Application . Analysis . Evaluation	Research Assignments - Teacher made. - Progressive throughout year.	. Difficult to assess these outcomes by objective tests - best measured by assignments. . These skills expected to develop over the year, hence assessment conducted over year. . Detailed marking guides provided to increase marker reliability.
Laboratory Skills	Laboratory Reports - Centrally set. - Progressive throughout year.	. Require the performance of basic laboratory operations and procedures. . These skills expected to develop progressively through the year.

IV. OUTLINING THE ASSESSMENT PROGRAM: COMPONENTS AND ALLOCATION OF MARKS

The culmination of the planning phase in test construction is an outline of the total assessment program showing forms of assessment to be used, assessment strategies to be followed and allocation and breakdown of marks.

The key to content validity at this stage of the planning lies in the allocation of marks. If content validity is to be maintained, the marks allocated to each component of the assessment program should accurately reflect the weightings or emphases given on the original table of specifications to the learning outcomes and topics tested by that component.

In addition, the various tasks, items and questions which make up the various assessment components should be distributed in such a way that they test a representative sample of the possible topics and learning outcomes which could be tested.

This may be achieved by observing the following series of steps, which will be illustrated by reference to the assessment program in Biology.

Step 1. Allocate a Weighting to Each Component of the Assessment Program.

During the previous procedure, the forms of assessment to be used in the program were determined and related to specific outcomes from the table of specifications (see Table 3.11). Now each of these forms of assessment (which are to comprise the components of the assessment program) should be given a weighting. This weighting should be the same as that originally given (on the table of specifications) to the outcomes assessed by these components.

This can be accomplished by drawing up a table similar to the one which follows which is based on the Table of Specifications for a Unit Course in Biology (Table 3.07).

Table 3.12 Allocation of Weightings for the Assessment Program in Biology.

FORMS OF ASSESSMENT	OUTCOMES TO BE ASSESSED	WEIGHTING
Objective Tests	Knowledge and comprehension	60%
Research Assignments	Higher Mental Abilities (Application, Analysis and Evaluation)	20%
Laboratory Reports	Laboratory Skills	20%
		100%

Note:

On the original table of specifications, for the unit course in Biology, Knowledge and Comprehension outcomes accounted for 60% of the total weighting. Therefore, objective tests will carry 60% of the total weighting for the assessment program. Similarly Research assignments assessing higher mental abilities will carry 20% and laboratory reports assessing laboratory skills will carry 20%.

Step 2. Breakdown the Weighting for Each Form of Assessment Across Subject Matter Divisions

The weighting assigned to each form of assessment (and related learning outcomes) can now be broken down into a smaller weighting for each subject matter division or topic tested by that form of assessment.

This can best be accomplished by constructing a table similar to the one which follows. It is very similar to the original table of specifications except that the components of the assessment scheme are now included and linked with outcomes and these are listed on the vertical axis for ease of preparation.

Table 3.13 Allocation of Weightings for Each Form of Assessment and Each Topic for the Assessment Program in Biology.

Topics / Form of Assessment (& Related Outcomes)	Cell	Nutr.	Resp.	Repr.	Ecol.	Total Weightings
Objective Tests (Knowledge and Comprehension)	20	15	9	9	7	60
Laboratory Reports (Laboratory Skills)	5	5	3	5	2	20
Research Assignments (Analysis, Evaluation Application)	5	5	3	4	3	20
Total Weightings	30	25	15	18	12	100

Note

Out of the total weighting of 60% for objective tests, the following weightings have been given to each topic: The Cell 20%; Nutrition 15%; Respiration 9%; Reproduction 9%; Ecology 2%; (Total 60%). These weightings for each topic correspond to the combined weighting for knowledge and comprehension given to each topic on the original table of specifications.

A similar check will reveal that the breakdown of the weightings for Laboratory Reports and Research assignments will also agree with the weightings originally assigned to the topics concerned.

Step 3. Work out a Mark Allocation for Each Component of the Assessment Program.

If the assessment program consists of one test worth 100 marks then no conversion from the previous type of table may be necessary. Each percentage weighting for each cell can be converted to the same mark allocation and the number of items within each cell is easily calculated.

If the assessment program is more comprehensive, however, the above arrangement may be too restrictive and may result in very small allocations of marks for individual questions and items.

This problem may be alleviated and more flexibility in mark allocations achieved by multiplying each weighting on the previous table by a constant. (See Table 3.14)

Table 3.14 Mark Allocations for Each Form of Assessment and Each Subject Division for the Assessment Program in Biology.

Topics / Forms of Assessment (and Related Outcomes)	Cell		Nutr.		Resp.		Repr.		Ecol.		Totals		Final Mark Alloc.
	%	M	%	M	%	M	%	M	%	M	%	M	
Objective Tests (Knowledge & Comprehension)	20	60	15	45	9	27	9	27	7	21	60	180	180
Laboratory Reports (Laboratory Skills)	5	15	5	15	3	9	5	15	2	6	20	60	60
Research Assignments (Application Analysis Evaluation)	5	15	5	15	3	9	4	12	3	9	20	60	60
Totals	30	90	25	75	15	45	18	54	12	36	100	300	300

Total Mark = 300

Note

The percentage weightings for each cell of the previous table have been multiplied by the constant three (3).

Step 4. Outline the Assessment Program.

As the culimination of the total planning procedure, the Assessment Program should be outlined in simple summary form showing:
a. The components of the Program; and
b. The mark allocations for each component.

In addition, to ensure that the links with the original table of specifications are maintained each component of the program should be broken down showing the number of component questions or tasks and mark allocations for each.

These final procedures are illustrated in the following tables.

Table 3.15 Assessment Program for Unit Course in Biology.

FORM OF ASSESSMENT	MARK ALLOCATION
One Centrally Set Objective Test (whole course)	90
Five Class Objective Tests (one for each topic)	90
Five Laboratory Reports (one for each topic)	60
Five Research Assignments (one for each topic)	60
Total	300 marks

Table 3.16 Breakdown of Marks for Each Component of Assessment Scheme in Biology

A. OBJECTIVE TESTS

One Centrally Set Test, 90 Questions, 1 mark each, as follows:

The Cell	30
Nutrition	22
Respiration	15
Reproduction	13
Ecology	10
	90

Five Class Tests, one on each topic, as follows:

The Cell	30
Nutrition	23
Respiration	12
Reproduction	14
Ecology	11
	90

B. LABORATORY REPORTS

Five Reports, centrally set, one on each topic, as follows:

The Cell	15
Nutrition	15
Respiration	9
Reproduction	15
Ecology	6
	60

C. RESEARCH ASSIGNMENTS

Five Assignments, teacher made, one on each topic, as follows:

The Cell	15
Nutrition	15
Respiration	9
Reproduction	12
Ecology	9
	60

EXERCISES

1. Prepare an outline of an assessment program based on your table of specifications. The outline should indicate the forms of assessment to be used and the related mark allocations.

2. Draw up a series of tables similar to tables 3.12 , 3.13 and 3.14 , showing how the mark allocations were arrived at and can be related back to the original table of specifications.

3. Breakdown the total mark for each component of the program showing how these marks are distributed across questions or items and across topics and learning outcomes.

REFERENCES AND FURTHER READING

BLOOM, B S, HASTINGS, J T and MADAUS, G F, Handbook on Formative and Summative Evaluation of Student Learning, McGraw-Hill, New York, 1971, Ch 23.

BROWN, F G, Principles of Educational and Psychological Testing, Dryden Press, 1970, Ch 10; Holt, Rinehart & Winston, 1976, Ch 13.

EBEL, R L, Essentials of Educational Measurement, Prentice-Hall, New Jersey, 1979, Ch 5.

GRONLUND, N E, Constructing Achievement Tests, Prentice-Hall, New Jersey, 1968, Ch 2; 1977, Ch 2.

THEOBOLD, J, Classroom Testing: Principles and Practice, Longman, Aust., 1974, Ch 1, pp 3-4; Ch 3, pp 28-30.

TOPIC 4

CONSTRUCTING TESTS AND TEST ITEMS

INTRODUCTION

The previous topic outlined and described the procedures to be followed in the planning of a valid and reliable assessment program.

The procedures stressed were:

1. Determine the purpose of the tests to be used.
2. Identify the subject matter and learning outcomes to be assessed on a table of specifications.
3. Determine appropriate forms of assessment and assessment strategies for the subject matter and outcomes being assessed.
4. Outline the assessment program showing components and mark allocations worked out so as to accurately reflect the importance of the topics and outcomes being assessed.

If these procedures are to lead to the desired result, however, that is an assessment program which is reliable and has high content validity, the forms of assessment to be used must be constructed properly and in such a way as to actually carry out the function which is planned for them. That is, they must actually test what they are intended to test.

This brings us to the next major procedure in the development of an assessment program: the construction of the actual tests and test items.

The principle to be followed here is that to ensure content validity, the tests and test items must be constructed in such a way as to actually bring out and assess the behaviours which they are constructed to assess.

Measures must be taken also, during the construction and scoring of the tests, to improve test reliability.

The treatment of this topic in this booklet shall be restricted to the consideration of three major types of test commonly employed in Technical and Further Education. These are:

1. Multiple-choice Tests.
2. Essay and Short Answer Tests.
3. Practical and Performance Tests.

The principles discussed, however, may be applied to the construction of other types of test, most of which are described in the references cited.

SPECIFIC OBJECTIVE

On completion of this topic, students will be able to construct multiple-choice, essay, short answer, practical and performance tests to measure the achievement of specified objectives.

1. MULTIPLE-CHOICE TESTS

a) What are Multiple-choice Tests?

Multiple-choice tests are paper and pencil tests which are made up of items which present several solutions or possible answers to a question or problem from which students have to select the correct one. Because they are objectively scored they are often referred to as objective tests.

The multiple-choice item generally consists of a stem which presents a problem and several alternative solutions to the problem or options. These options include one correct answer (the key) and several plausible incorrect answers called distractors.

Example:
The structural steel section shown in the cross sectional diagram is a: — stem

* a. universal beam; — key
b. universal column;
c. rolled steel joist; — distractors
d. channel.

options

The stem of a multiple-choice item may be written in the form of a statement as in the above example or in the form of a question.

Example:
What is the maximum permissible length of 15mm diameter copper pipe to supply water to a bath and shower?

* a. 2135mm;
b. 6000mm;
c. 600mm;
d. no limit.

The function of the distractors is to distract those students who are uncertain of the correct answer. Generally 4 or 5 options are employed.

Multiple-choice items are often improved by the addition of a diagram to add realism.

The examples shown above are referred to as the single or simple item type. More complex items may also be constructed and these are discussed by Gronlund (1977, pp 70-73)

b) When are Multiple-choice Items Used?

Multiple-choice items are commonly used in TAFE to measure the basic cognitive learning outcomes such as knowledge, comprehension and application. The higher order learning outcomes such as analysis, synthesis and evaluation can also be assessed effectively by multiple-choice items but such items are very difficult to construct and are beyond the scope of this booklet. (See Gronlund, 1977, Ch 4)

Multiple-choice tests are also extremely useful when a broad range of theoretical content is to be sampled in any one test or test program such as a final test of Trade Theory. They are useful also for progressive tests, internal teacher made tests, centrally set tests, general achievement or mastery tests.

In fact, Gronlund goes so far as to say that:

> when the various test items are equally adaptable to the learning outcomes and subject matter to be measured, multiple-choice items will generally provide a more adequate measure than the other item types.
>
> (Gronlund 1977, p35)

This viewpoint, shared by many experts on test construction, is brought about because multiple-choice tests enjoy a number of advantages over other paper and pencil tests, such as essay and short answer tests, in relation to both reliability and content validity. For instance:

1. They can cover a wide and therefore representative sample of course content.
2. They establish tight control over content and outcomes being assessed.
3. They are objectively and therefore reliably scored.
4. They are less affected by factors such as writing ability, expression, bluffing, etc.

It should be pointed out, however, that these advantages rely very heavily on good item construction which can be both time consuming and difficult. Not only are multiple-choice items subject to chance factors such as lucky guesses, but it is common to find items which are ambiguous, too easy, too difficult, lacking in discrimination power or which fail to measure the important and worthwhile learning outcomes.

All of these factors, of course, may adversely affect the validity and reliability of the test (See Topic 2).

If multiple-choice items are to be used in an assessment program, therefore, and if they are to do the job intended for them, a number of general rules or guidelines for their construction must be adhered to.

c) Guidelines for Constructing Multiple-choice Items

> Ideally, a multiple-choice item presents students with a task that is both important and clearly understood, and one that can be answered correctly only by those who have achieved the desired learning.
>
> (Gronlund, 1977, p39)

To achieve this ideal, Gronlund provides a number of guidelines or general rules of item construction which are outlined below:

1. Design each item to measure an important learning outcome.
2. Present a single clearly formulated problem in the stem of the item.
3. State the stem of the item in simple clear language.
4. Put as much of the wording as possible in the stem of the item.
5. State the stem of the item in positive form wherever possible.

6. Emphasise negative wording whenever it is used in the stem of an item.

7. Make sure that the intended answer is correct or clearly best.

8. Make all alternatives grammatically consistent with the stem of the item and parallel in form.

9. Avoid verbal cues that might enable students to select the correct answer or to eliminate an incorrect alternative.

10. Make the distractors plausible and attractive to the uninformed.

11. Vary the relative length of the correct answer to eliminate length as a clue.

12. Avoid using the alternatives 'all of the above', 'none of the above', 'I'd rather not say' or similar.

13. Vary the position of the correct answer in a random manner.

14. Control the difficulty of the item either by varying the problem in the stem or by changing the alternative.

15. Make certain each item is independent of the other items in the test.

16. Use an efficient item format.

(Gronlund, 1977, pp39-53)

These general rules or guidelines are fully explained and illustrated by Gronlund and can be easily understood by reference to that text. If adhered to, they will do much to ensure the reliability and validity of multiple-choice tests.

In addition, however, multiple-choice items must be constructed so as to bring out and assess the specific outcomes they are intended to measure. This requires specific procedures for the assessment of different outcomes. These will now be considered in relation to the cognitive outcomes: knowledge, comprehension and application.

d) Multiple-choice Tests of Knowledge

Multiple-choice tests of knowledge require students to recall material in essentially the same form as it was learned.

It will be recalled that the achievement of knowledge outcomes usually requires students to demonstrate that they can recall or recognise facts or specific aspects of knowledge. Multiple-choice tests of knowledge can readily and easily be constructed therefore by following the general rules of item construction outlined above and by requiring students to recognise a correct statement (from alternative incorrect statements) which is presented in precisely the same form as it was presented in a text book or in class.

In this way, multiple-choice items may be constructed which are ideally suited to measure knowledge of the wide range of specific facts, terminology, conventions, methods, classifications, categories, rules, principles, etc commonly taught in most TAFE courses.

Examples:

i) The most suitable method of joining aluminium sheeting and tube is by means of:
   a) fusion welding;
   b) brazing;
* c) argon arc welding;
   d) soft soldering. ( )
   (knowledge of methodology)

ii) The approximate temperature of an oxy acetylene welding flame is:
* a) 3000 C;
   b) 3000 F;
   c) 6000 F;
   d) 6000 C. ( )
   (knowledge of specific facts)

iii) All pipes and fittings used in a water service must be capable of withstanding a hydrostatic pressure of:
   a) 150lbs per square inch;
   b) 200lbs per square inch;
* c) 300lbs per square inch;
   d) 400lbs per square inch. ( )
   (knowledge of convention)

iv) A 'direct connection' is one which:
   a) maybe used if required;
* b) takes the shortest distance to the property served;
   c) makes possible backflow of polluted water to the main;
   d) has an air gap. ( )
   (knowledge of terminology)

The above examples all constitute tests of direct or factual knowledge because they all measure recall of factual information presented in the same form as it was taught.

Such items are comparatively easy to write and are effective as long the outcomes being assessed do not go beyond direct knowledge.

Unfortunately, because they are easy to write, multiple-choice items testing knowledge will often predominate in a test at the expense of more complex outcomes which should be tested. This, of course, adversely affects the content validity of the test.

In order to construct multiple-choice items testing more complex outcomes than direct knowledge, such as comprehension and application, two further guidelines must be followed. These are:

1. An element of novelty must be introduced into the item. That is, the problem or question must be presented in a form and in language different to the way it was presented in class.

2. The item must be constructed in such a way as to require the student to demonstrate that he has attached meaning or understanding to his knowledge (comprehension) and/or that he can use it effectively in dealing with a novel or problem situation. (application)

e) Multiple-choice Tests of Comprehension

Multiple-choice tests of comprehension require students to demonstrate that they have understood or can attach meaning to facts, information, concepts, principles, etc.

It will be recalled that comprehension is usually tested by requiring students to do the following:

1. give the essential meaning of a concept or idea in other words (interpretation);

2. explain the reason for a particular event, phenomenon or relationship (explanation); or

3. predict the outcome of an event, phenomenon or relationship (prediction).

Comprehension may be tested by multiple-choice items, therefore, in the following ways:

1. Item requires student to recognise the essential meaning of a concept or idea (interpretation)

   Example:
   In the processing of milk for sale, Vitamin D concentrate has been added to provide at least 400 International Units per litre. The carton can then legally state that the milk is:

   * a) fortified;
   b) processed;
   c) pasteurised;
   d) refined. (Multiple-choice Test Manual, p21)

   This item requires the student to comprehend the essential meaning of the concept: 'Fortified' although the account provided in the question is different to the formal definition provided in class.

2. Item requires the student to recognise the correct reason for a phenomenon, event or relationship (explanation)

   Example:
   When welding an inside corner joint, a flame adjustment of a slightly oxidising nature is used because:

   a) the amount of penetration should be kept at a minimum;
   b) the inner cone has a tendency to touch the metal;
   c) it provides for the unusual amount of contraction;
   * d) it is difficult to obtain additional oxygen from the air.

   (Bloom et al, 1971, Ch 23)

3. Item requires the student to recognise the correct outcome of a proposed event (prediction)

   Example:
   If an electric refrigerator is operated with the door open in a perfectly insulated sealed room, what will happen to the temperature of the room?

* a) it will rise slowly;
b) it will remain constant;
c) it will drop slowly;
d) it will drop rapidly.

(Multiple-choice Test Manual, p25)

This item requires the student to correctly predict the outcome of an action which in turn requires an effective comprehension of the underlying principles of refrigeration.

f) Multiple-choice Tests of Application

Multiple-choice tests of application also measure understanding but they go beyond this to the application of that understanding. Not only must the student demonstrate that he comprehends some aspect of knowledge, but application items must require him to apply or use that knowledge in coping with a novel or problem situation.

Application outcomes are normally tested by requiring students to describe the application of some aspect of knowledge - fact, principle, rule, method, etc - in a given situation; or by requiring the student to actually apply the knowledge to the solution of a specified problem.

Application may be tested by multiple-choice items in the following ways:

1. Item requires student to recognise the correct application of a fact, rule, principle, method, etc in a given situation

   Example:
   Analysis of the responses to a multiple-choice item reveals that only 2 students out of 100 selected the fourth distractor. Which would be the best change to make to that item?

   a) change the stem to incomplete statement form;
   b) eliminate the fourth distractor;
   * c) make the distractor more plausible;
   d) vary the position of the distractor.

   This item requires the student to apply his knowledge of the guidelines for constructing multiple-choice items outlined earlier.

2. Item requires student to recognise the correct solution to a given problem

   Example:
   A milling cutter with 6 teeth is turning at 130rpm. For a chip load of 0.004, the correct feed in inches per minute would be:

   a) 0.312;
   * b) 3.12;
   c) 6.24;
   d) 9.36.

(Bloom et al, 1971, Ch 23)

Exercises

1. Examine the multiple-choice items in Appendix 1. What outcomes do they measure?

2. Change these items so that they measure more complex learning outcomes.

3. Select two topics from your assessment program planned in Topic 3. For each topic write a multiple-choice item testing: a) knowledge; b) comprehension; c) application.

II ESSAY AND SHORT ANSWER TESTS

a) What are Essay and Short Answer Tests?

Essay and short answer tests are paper and pencil tests. They are coupled here because they both require students to supply rather than select the correct answer. They also have a number of other features and construction procedures in common. The chief difference between the two types of test is found in the type and extent of response to be supplied.

Essay tests require students to produce an essay in answer to a question. They provide a freedom of response required in measuring certain complex outcomes. The amount of freedom provided depends on the nature of the question and practical considerations such as time available, etc.

Restricted-Response essay questions place strict restrictions on the scope and form of the essay to be written by clearly defining the nature of the answer required (using terms such as 'define', 'explain', 'descibe', etc) and by clearly indicating the scope of answer required (eg, 'two advantages', 'two disadvantages', etc).

Extended-Response essay questions give the student much more freedom in determining and organising the scope and form of the answer for himself. (For examples of extended-response and restricted-response essay questions, see Appendix 1).

Whichever form is used, however, the student must be given enough freedom and enough guidance to ensure that the desired behaviours can be produced and assessed.

Short answer tests restrict the answer to be supplied even further - to a paragraph, sentence, phrase or missing word (see Appendix 1). The latter are sometimes called 'completion' items. They may also involve calculations (where a number or numbers are required), and may be accompanied by diagrams to add realism.

Jones and Whittaker provide a useful example of the realistic short answer item consisting of four photographs of motor car components and the candidate is required to carry out the following operations with regard to each component:

i) identify the component;
ii) describe the defect;
iii) describe the probable cause;
iv) suggest remedial action.

(Jones & Whittaker, 1975, pp46-47)

Short answer items must be constructed in such a way as to produce answers which can be clearly identified as correct or incorrect.

b) When are Essay and Short Answer Tests Used?

Essay tests are commonly used in TAFE to measure the higher cognitive outcomes such as analysis, synthesis and evaluation. Gronlund suggests that essay tests are

used to measure certain complex outcomes which include: 'the ability to create, to organise, to integrate, to express and similar behaviours that call for the production and synthesis of ideas.' (Gronlund, 1977, p75).

These outcomes are more commonly found in Technician level courses in TAFE than in Trade courses.

In addition, Ebel suggests that essay tests may also be used for the exploration and assessment of student attitudes. (Ebel, 1979, p101)

In this way, essay tests enjoy an advantage over multiple-choice and other objective tests which are not so well suited for measuring these behaviours.

Essay tests may be used in final or progressive tests and in teacher made or centrally set tests. They are more useful in general achievement than in mastery tests.

Short answer tests are used (like multiple-choice tests) chiefly to measure the basic cognitive outcomes; recall of factual knowledge; comprehension and application.

They are used widely in this way in the assessment of Trade courses. They are less influenced by guessing than multiple-choice items and many teachers prefer them because they require students to recall, explain, or work out the correct answers for themselves rather than recognise them.

Short answer tests may be used in final or progressive tests and in teacher made or centrally set tests. They tend to be more useful than essay tests when testing for mastery.

Unfortunately, however, both essay and short answer tests suffer from some very serious shortcomings, which must be recognised and combatted if they are to be used effectively in an assessment program.

These include the following:

1. Essay questions are time consuming to answer so that essay tests tend to be restricted in the scope and sample of achievement which is assessed. This problem need not be so severe for short answer tests as long as irrelevant behaviours such as sketching are eliminated.

2. Essay tests tend to be over influenced by such factors as writing ability, punctuation, grammar, bluffing, etc (see Gronlund, 1977, p76).

3. Both essay and short answer tests lack control over the content and outcome assessed. It is very difficult to write questions which eliminate all rival interpretations and ensure that the student produces the response which the examiner wants to assess. (See examples in Appendix 1).

4. Both essay and short answer tests are subjectively scored. Consequently, numerous studies have demonstrated that the scoring of these tests can be inconsistent from marker to marker and from time to time (see Gronlund, 1977, p77). This shortcoming, referred to as marker unreliability, is a well known shortcoming of essay tests. It can also have an adverse effect, however, on the marking of short answer and even 'completion' items.

All of these shortcomings may adversely affect the reliability and validity of a test. It is because of these shortcomings that many test constructors suggest that essay and short answer tests should be avoided altogether and recommend the use of objective tests such as multiple-choice tests whenever possible (see Jones & Whittaker, 1975, p45).

If essay and short answer tests are to be used, therefore, to measure complex outcomes or for any of the reasons outlined above a great deal of care must be taken both with their construction and in their scoring to minimise the impact of their inherent shortcomings.

With this in mind both Gronlund (1977) and Ebel (1979) provide a number of suggestions for the preparation and scoring of essay and short answer tests. These are outlined in the following sections.

c) Suggestions for Preparing Essay Tests

1. Identify very clearly the outcome or outcomes to be assessed.
2. Relate the questions as directly as possible to the learning outcomes being measured.
3. Formulate questions that present clear tasks to the student.
4. Provide ample time for answering the questions and suggest a time limit on each question.
5. Ask questions that are determinate (ie, experts could agree that one answer is better than another).
6. Define the examinee's task as completely and specifically as possible.
7. In general, give preference to more specific questions that can be answered more briefly.
8. Avoid giving the examinee a choice among optional questions unless special circumstances make such options necessary.
9. Test each question by writing an ideal answer to it.

(Adapted from Gronlund, 1977, pp80-83; Ebel, 1979, pp102-104)

d) Suggestions for Writing Short Answer Items

1. Word the question or incomplete statement carefully enough to require a single, unique answer.
2. Think of the intended answer first. Then write a question to which that answer is the only appropriate response.
3. If the item is an incomplete sentence try to word it so that the blank comes at the end of the sentence.
4. Use a direct question unless the incomplete sentence permits a more concise or clearly defined correct answer.
5. Avoid unintended clues to the correct answer.

6. Word the item as concisely as possible without losing specificity of response.

7. Arrange space for recording answers on the right margin of the question page.

8. Avoid using the conventional wording of an important idea as the basis for a short answer item.

(Ebel, 1979, pp166-9)

e) Suggestions for Scoring Essay and Short Answer Tests

1. Evaluate answers in terms of the learning outcomes being measured.

2. Score restricted response and short answer questions by the point method using a model answer or scoring key as a guide.

3. Grade extended response answers by the rating method using defined criteria as a guide.

4. Evaluate all of the students' answers to one question before proceeding to the next question.

5. Evaluate answers without knowing the identity of the writer.

6. Wherever possible have two or more persons grade each answer.

(Adapted from Gronlund, 1977, pp83-86)

All of these procedures are fully explained and illustrated in the references cited.

It should be stressed that if essay or short answer questions are to be scored using the point or analytical method of scoring (as a means of improving marker reliability) then a carefully prepared scoring key or marking guide should be provided.

This guide should contain the following minimum components:

i) an analysis of the desired or model answer into basic or minimum components, steps or points;

ii) a mark allocation for each of the above divisions;

iii) clear objective instructions on the procedure for arriving at these marks (ie, how marks are to be given or deducted).

(For further guidance on this, see Ebel, 1979, p96 and pp104-106)

Exercises

1. Examine the essay and short answer questions in Appendix 1. Which items exhibit the shortcomings dexcribed on page 50? What measures might be taken to improve these items?

2. Select a suitable topic from your assessment program planned in Topic 3. Write two essay questions and two short answer questions to measure nominated learning outcomes related to this topic.

3. Provide marking guides or scoring keys for each question.

III PRACTICAL AND PERFORMANCE TESTS

a) What are Practical and Performance Tests?

Practical and performance tests are those tests which require the student to produce some aspect of job or practical performance. Practical tests are usually those associated with trade practical courses whereas the term 'performance test' is often used to cover any aspect of job performance. Often the two terms are used synonymously.

It will be recalled that Jones and Whittaker use the term 'Hands-on test' to describe this type of test and suggest two types as the most appropriate for testing industrial skills. These are the Work Sample and the Skill Sample tests. (Jones and Whittaker, 1975, pp36-41; and see Appendices 2 and 3)

The skill sample test will often require the construction of a product such as a taper or vee pulley, whereas the work sample test may require the performance of a task or process, eg, 'removal and replacement of a clutch plate'.

Gronlund provides a four way classification of performance tests: (i) paper and pencil performance; (ii) identification; (iii) simulated performance; and (iv) work sample. (1977, pp90-94)

The chief distinguishing feature of all of these performance or 'hands-on' tests is that they require students to actually perform some process or procedure in a realistic or job-like situation with tools, materials or equipment which they would be likely to use on the job. This is so whether the process or procedure requires performance with paper and pencil or performance with a lathe or word processor. It is this which distinguishes them from the paper and pencil tests so far discussed which require students to demonstrate that they know the process or procedure (knowledge); that they can explain its operational principles (comprehension); or that they can describe the method of operation (application).

As Gronlund puts it:

> if you want to determine what an individual 'knows about' a given performance, a knowledge test is perfectly appropriate. However, if you want to describe his 'proficiency in performing an activity', a performance test must be used'.
>
> (Gronlund, 1977, p88)

b) When are Practical and Performance Tests Used?

Practical tests are commonly used to measure the wide range of practical or psychomotor skills taught in Trade practical courses. They include work sample tests which require the performance of a discrete job or task associated with an occupation and skill sample tests which require the performance of skills and operations associated with these jobs and tasks.

These latter are sometimes referred to as exercises and measure such trade skills as filing, cutting, measuring, screw cutting, taper turning, etc.

Performance tests can be used to measure the wide range of job or occupational skills taught in other TAFE courses. These range from the more academic skills taught in Commerce and Management courses for instance, to laboratory type skills taught in Science, Applied Biology, Agriculture and Horticulture courses and basic occupational skills such as the typing of a business letter or the preparation of a dinner party taught in Secretarial and Home Economics courses. Many of these skills have manual as well as mental or cognitive elements.

Practical and performance tests are suitable for final or progressive tests, internal or centrally set tests. They are particularly suitable for mastery or criterion-referenced testing as precise conditions and performance criteria can be established which determine which individuals have mastered specified aspects of performance.

They are useful also for measuring certain outcomes in the affective domain of educational endeavour - outcomes such as the development of conscientious attitudes and safe habits of work, etc.

Despite these uses and potential advantages, however, practical and performance tests suffer from some of the same shortcomings and problems as essay tests. For instance:

1. They are often time consuming to perform and therefore restrictive in the scope of performance which can be sampled.

2. They are often ambiguous and unclear in relation to the task to be performed and the type and standard of performance required.

3. They are just as prone to the hazards of subjective and unreliable marking as essay tests.

These shortcomings, as noted earlier, can adversely affect both the validity and reliability of a test. To avoid the pitfalls which they represent or to minimise their effects, a great deal of care must be taken in construction and scoring.

Many of the suggestions for the preparation and scoring of practical and performance tests are similar to those already outlined for the preparation and scoring of essay tests.

c) Suggestions for Preparing Practical and Performance Tests

1. Select jobs and tasks so that an adequate and representative sample of performance is measured.

2. Select jobs and tasks, and design each test so that the critical performance is produced and assessed.

3. Define the student's task as completely and specifically as possible. This may involve the preparation of a job sheet which contains:

   - job title;
   - objectives of task or job;
   - steps of procedure;
   - tools and equipment required;
   - diagram showing specifications, materials, tolerances, etc (see examples in Appendices 2 and 3)

4. Provide adequate and clear guidance on standards required and criteria to be used in marking.

5. Provide ample time for completion of the test and indicate whether time for completion is an assessment factor.

6. Work through the test yourself to check that guidance, materials, time allowed, etc are adequate and that the test actually requires the type of performance to be assessed.

d) Suggestions for Marking Practical and Performance Tests

1. Evaluate the product or the process in terms of the performance being assessed and the criteria put forward.

2. Where possible score the product or the process by the point method using a marking key as a guide.

3. Evaluate performance without knowing the identity of the examinee where this is possible.

4. Use more than one independent marker if possible.

In relation to the construction of a marking guide or scoring key for practical and performance tests, the following suggestions might be followed.

1. Where the test requires the construction of a finished product, such as a Vee Pulley, for example, analyse the product into component parts or sections and allocate marks for each part. Some aspect of performance such as safety, speed, marking out, etc may also be allocated marks and marks allocated should reflect the importance or prevalence of the skills involved. (See marking guide in Appendix 4)

2. Where the test requires the performance of a task or process (removal, replacement, adjustment, etc) analyse the task into component steps or operations and allocate marks for each step or operation. Some aspect of the completed task (eg, finish) may also be allocated marks and marks allocated should again reflect the importance or prevalence of the skills involved. (see marking guide in Appendix 2)

3. Provide clear and unambiguous criteria for the allocation and/or deduction of marks. (See marking guide in Appendix 4)

It should be pointed out also that whereas the above suggestions are useful general guidelines, the precise form of the marking guide must vary depending on the nature of the task or test.

NOTE

1. Many of the suggestions and guidelines put forward regarding the construction and scoring of essay, short answer, practical and performance tests may be applied to the construction and scoring of assignments, projects and reports as these often contain essay, short answer, practical and performance elements.

2. When the tests, test items and other components of the assessment program have been constructed, the tests have to be reviewed, edited and arranged in a form suitable for administration. Directions for supervisors and examinees should also be written and included. (These final steps are described simply and adequately in Gronlund, 1977, Ch 7)

EXERCISES

1. Examine the practical tests and marking guides provided in Appendices 2, 3 and 4. In what ways might they be improved.

2. Select a topic from your assessment program which has practical or performance outcomes. Design a practical or performance test for this topic, giving full details on task or job, objectives, steps, tools and equipment, specifications, assessment criteria, etc.

3. Design a marking guide for this test which shows:

   a) breakdown of total task into elements;
   b) distribution of marks for each element;
   c) criteria for allocation of marks.

REFERENCES

BLOOM, B S, HASTINGS, J T and MADAUS, G F. Handbook on Formative and Summative Evaluation of Student Learning. McGraw-Hill, New York, 1971, Ch 23.

BROWN, F G. Principles of Educational and Psychological Testing. Dryden Press, 1970, Ch 10; Holt, Rinehart & Winston, 1976, Ch 13.

EBEL, R L. Essentials of Educational Measurement. Prentice-Hall, New Jersey, 1972, Chs 6,7 and 8; 1979,Chs 6,7,8 and 9.

GRONLUND, N E. Constructing Achievement Tests. Prentice-Hall, New Jersey, 1968, Chs 3,4,5 and 6; 1977, Chs 3,4,5,6 and 7.

JONES, A and WHITTAKER, P. Testing Industrial Skills. Gower Press, UK, 1975, Ch 4.

Multiple-Choice Test Manual: A Guide for Teachers. NSW Department of TAFE, Division of Research and Development.

THEOBOLD, J. Classroom Testing: Principles and Practice, Longman, Aust, 1974, Ch 2.

TOPIC 5

TREATMENT OF TEST RESULTS

INTRODUCTION

The previous topic described the final procedures to be followed in the development of the actual assessment program.

No account of educational measurement would be complete, however, and no attempt to develop an assessment program should be finalised without some consideration of the proposed treatment of test results.

It will be recalled that achievement test results are used by teachers and educators for three main functions:

1. To rank or grade students in terms of their relative achievement of course objectives.

2. To contribute evaluative data for instructional decision making.

3. To evaluate the effectiveness of the tests employed.

In relation to the third of these functions, Gronlund provides a simplified item analysis procedure to evaluate the effectiveness of crition-referenced and norm-referenced items. (1977, pp110-116). The wider use of test results for evaluative purposes, however, is almost a course in itself and is well beyond the scope of this booklet.

Consequently the treatment of this topic in this booklet shall be restricted to a brief consideration of the ways in which test scores are used to rank or grade learners. That is, the ways in which achievement test scores are converted to marks or grades.

This will be considered in relation to two different approaches to test interpretation applied to two different types of test.

1. Criterion - Referenced Testing (the mastery test)
2. Norm-Referenced Testing (the general achievement test)

SPECIFIC OBJECTIVE

On completion of this topic, students will be able to convert achievement test scores to grades.

I CRITERION - REFERENCED TESTING

It will be recalled that when this approach to testing is employed, test scores are interpreted or measured against a set standard or criterion. Grades, which are usually pass or fail, reflect achievement or non-achievement of the criterion. This form of testing is usually associated with mastery tests which are used to measure absolute achievement or mastery of a narrow range of objectives.

These tests usually consist of items, tasks or sets of items designed to measure specified objectives and results in the test are organised and interpreted in relation to the mastery of these objectives.

Consequently the procedure for converting scores to grades is a simple one as follows:

1. Determine a performance standard or acceptable level of mastery for each individual objective, task or sub-test and for the total test or assessment program.

   This standard might be in terms of:

   a) speed of performance;

   b) accuracy of performance;

   c) percentage of items correct or marks scored. (eg, 80% of terms correctly recalled, 80% score on total program, 8/10 tests passed of total program)

   (Adapted from Gronlund, 1977, p118)

2. Determine marks or raw scores achieved on each objective, task or sub-test and for the total test or program. (according to procedures specified on the tests and marking guides - see Topic 4)

3. Allocate the grade PASS for each score which attains or surpasses the prescribed level for mastery and FAIL for scores which do not. (Grades of pass may also be used if discrimination within the pass category is required - these procedures are described in the next section).

   This will usually involve allocating pass or fail grades for each objective, task or sub-test as well as for the total test or test program.

Gronlund describes a useful method of reporting results on a mastery test based on a careful analysis of performance in relation to each objective. This not only represents a more meaningful reporting of results than a simple overall PASS or FAIL, but is very useful for remedial follow up which is generally associated with the mastery approach. (Gronlund, 1977, pp118-119)

II NORM-REFERENCED TESTING

It will be recalled that when this approach to testing is employed, test scores are interpreted, not so much in relation to a set standard or criterion, but in relation to standards within the group or groups of students tested. This approach to testing is usually associated with general achievement tests which are used to measure general or relative achievement over a broad range or sample of course objectives, (rather than each specific objective). Students are ranked or compared in their achievement of these objectives by the use of grades such as A B C D E etc or DISTINCTION, CREDIT, PASS, FAIL.

A simple method for the allocation of these grades using raw or untreated scores will now be described.

1. Determine raw scores for each student on each component of the assessment program - according to procedures set down in the marking guides.

   According to Brown, the eventual grades will have more validity if these scores are based on a variety of assessment procedures and if representative samples of subject matter and outcomes are tested (Brown, 1976, p268).

2. Combine raw scores from each component into a total or combined score for the assessment program - in the module, unit or subject.

   This can only be done if each of the component scores is expressed on a similar quantitative scale (eg, mark out of 100). If this is so and the distributions on these component variables are roughly comparable, Brown maintains that the relative rankings obtained by combining raw scores in this way will not differ greatly from those obtained by combining standard or derived scores. (Brown, 1976, p268)

   Scores from particular components may of course be given increased weightings. (For a discussion of this and other aspects of combining class marks, final marks, etc in TAFE - see Hermann, 1969, pp31-35).

3. Rank the total scores from high to low and make a frequency distribution showing the number of students achieving each score. (See Gronlund, 1977, p122)

   This procedure readily shows the position or standing of each student within the group. For example, a score of 37/40 on Gronlund's distribution (1977, p122) would place that student in the top 10% of students who attempted the test - or better than 90% of the students. The score itself does not determine the grade - it is the standing of that score in relation to the other scores.

4. Allocate grades according to some previously determined distribution or method.

   Grades are allocated for each score according to the position of that score in the distribution. For example, the top 10% might be given A's, the next 10% or 20% might be given B's and so on. Some aspects of an absolute scale, such as a score of 50% to pass might also be employed.

   'Probably the most common procedure', says Brown, 'is to have certain expected percentages in mind, but apply these percentages flexibly depending on the performance of the particular class.' (1976,p268). Some subjective element, however, will always be involved.

Sometimes, however, the examiner may wish to compare a student's performance on different tests. Sometimes, also, he may want to compare or combine test scores which are expressed on a quantitative scale with quite different measures of achievement expressed as qualitative scores or ratings, which are not comparable.

In these and other circumstances the examiner may have to convert the raw scores to so called standard scores. This allows comparison of scores from different distributions because standard scores have a common mean and standard deviation.

To understand this process, however, it is necessary to introduce basic descriptive statistics such as the mean and standard deviation which are beyond the scope of this booklet. (For further reading on this see Gronlund, 1977, pp120-128; Brown, 1976, pp180-190).

REFERENCES AND FURTHER READING

BROWN, F G. Principles of Educational and Psychological Testing. Holt, Rinehart & Winston, 1976, Ch 10 and Ch 13, pp267-8.

EBEL, R L. Essentials of Educational Measurement. Prentice-Hall, New Jersey, 1972, Chs 11 and 12; 1979, Chs 10 and 11.

GRONLUND, N E. Constructing Achievement Tests. Prentice-Hall, New Jersey, 1968, Ch 7; 1977, Ch 8.

HERMANN, G. 'Examining Systems in Technical Education'. The Australian Technical Teacher. May 1969.

THEOBOLD, J. Classroom Testing: Principles and Practice, Longman, Aust, 1974, Chs 3, 4 and 6.

APPENDICES

Appendix 1. Paper and Pencil Tests.

Appendix 2. A Skill Sample Test: Construction of a Vee Pulley.

Appendix 3. Marking Guide for a Skill Sample Test: Construction of a Soldering Iron.

Appendix 4. The Taxonomy of Educational Objectives: Summary of the Three Domains (see Booklet: Curriculum Construction in TAFE; Appendix 1).

Appendix 5. Behavioural Terms: Some examples of 'action verbs' in the three domains (see Booklet: Curriculum Construction in TAFE; Appendix 2).

APPENDIX 1

PAPER AND PENCIL TESTS

A. Supply Type

1. Essay - Extended Response

Describe the assessment procedures which you would use to measure achievement in a course which you teach. Justify these procedures in terms of reliability and validity and explain the measures you would adopt to ensure that the assessment program had high content validity. (1500 words)

2. Essay - Restricted Response

Describe two advantages and two disadvantages of the restricted response essay question. (one page)

3. Short Answer

i Briefly explain the following:

a. cost centres;
b. direct costs;
c. prime costs;
d. flexible budget. (half page for each)

4. Completion

The tool shown in the diagram is a _____ chisel.

B. SELECTION TYPE

1. True - False

i Hard drawn copper tube may be joined using soft soldered capillary fittings. T F

ii Where only low pressure water supply is available, the mitered elbow fitting would give the best flow. T F

2. Matching

Match the tools listed in Column 1 with the uses shown in column 2.

Column 1	Column 2
Flat cold chisel	sheering
jenny calipers	marking
solder pot	tinning
etc.	etc.

3. Multiple Choice

Indicate which is the correct answer by writing its corresponding letter in the margin.

i Where only low pressure water supply is available, which of the following fittings would give the best flow?

a. elbows;
* b. bends;
c. mitered elbows;
d. M and F elbows. ( )

ii The chisel shown in the diagram is a:

a. diamont point;
* b. cross cut;
c. round nose;
d. flat. ( )

APPENDIX 2

A SKILL SAMPLE TEST: Construction of a Vee Pulley*

Aim: To turn an angular groove and produce a closely toleranced bore in aluminium.

Material: 60 x 32 Aluminium.

Safety: Keep hands clear of revolving parts.

Procedure: (Demonstrations will be given by the teacher.)

1. Set aluminium true in a 4-jaw chuck.
2. Face end.
3. Turn 58 and 36 diameters.
4. Turn vee.
5. Drill and finish the hole.
6. Cut the keyway.
7. Bevel, deburr and part-off.
8. Hold lightly by 36 diameter and skim parted face if necessary.

Feature	Requirements
Diameters	Concentric and within limits of size.
Lengths	Within limits.
Hole	Bore and keyway correct shape, size.
Vee	Correct shape and size.
Finish	As specified. No burrs or sharp edges.

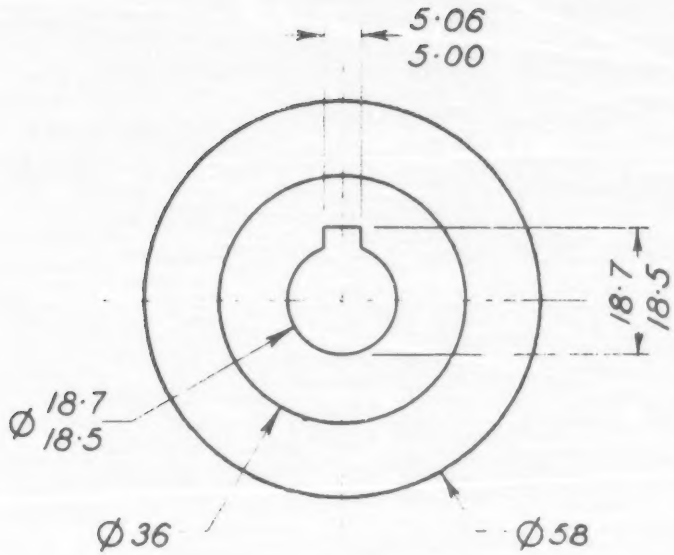
Issued subject to alteration pending determination of S1 surface finish symbols.

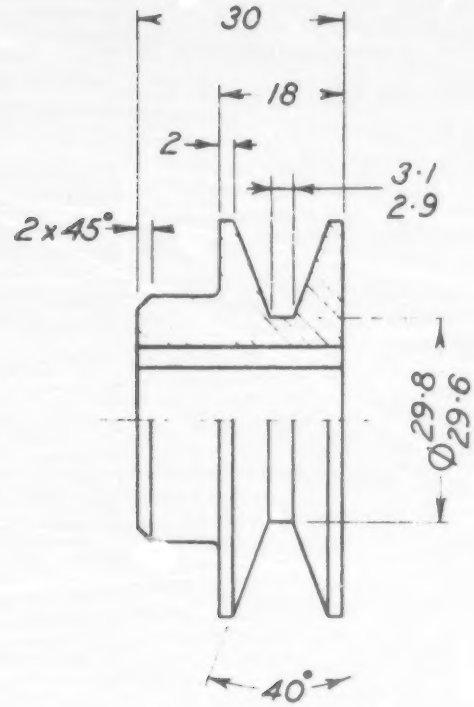
NOTE:
The student's interest, attitude and diligence towards his work, and towards the teacher's demonstrations MUST be considered in the assessment of this exercise.

* Test supplied by NSW Department of TAFE, School of Engineering Trades.

UNLESS STATED OTHERWISE DIMENSIONS ARE IN MILLIMETRES.

3RD ANGLE PROJECTION





Tolerance : ± 0·2 unless shown otherwise.

Finish : ▽ all over

Surface finish symbols are expressed in micrometres

MATERIAL : Ø 60 x 32 Aluminium.

APPENDIX 3

MARKING GUIDE FOR A SKILL SAMPLE TEST:

Construction of a Soldering Iron*

A. Distribution of Marks

MARKS FOR SOLDERING IRON

Overall Bit Length	- 5
Handle Length	- 5
Flat Filing	- 25
Handle Square	- 5
Bit Square	- 25
Surface Finish	- 5
Angles	- 20
Handle Centre	- 5
Handle Finish & Number	- 5
Total Marks	100

* Courtesy AIS Apprentice Training School.

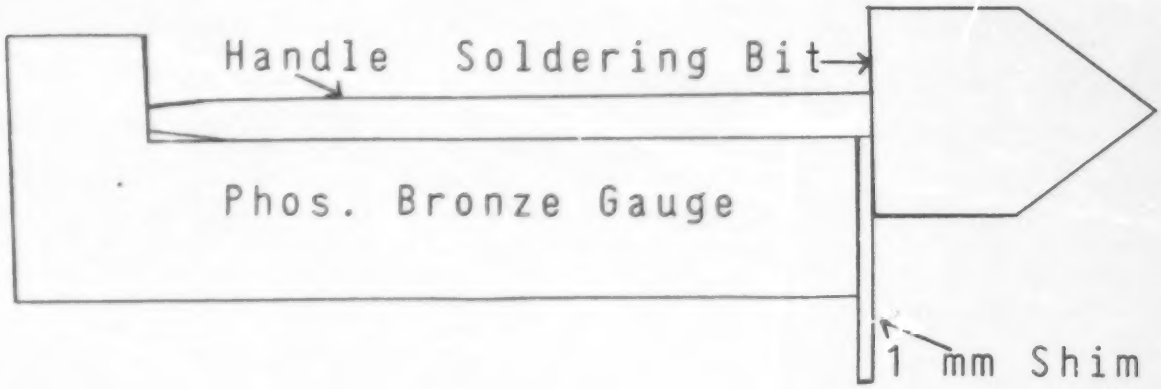
B. Criteria for Allocation of Marks

OVERALL BIT LENGTH

Using Vernier Calipers award 5 marks for length between 70 mm - 72 mm. Deduct 2 marks for each 1 mm over or under.

HANDLE LENGTH

Using Gauges (No. 1) award 5 marks up to 1 mm under or over.



FLAT FILING

Check on surface plate dependant on blued surface. 5 marks per side.

HANDLE SQUARE

Using Square. Visual check.

BIT SQUARE

1 Award 5 marks for each face by placing square across largest width and with distance marked off square into 5 equal lengths. Allow 1 mark for each square touching.

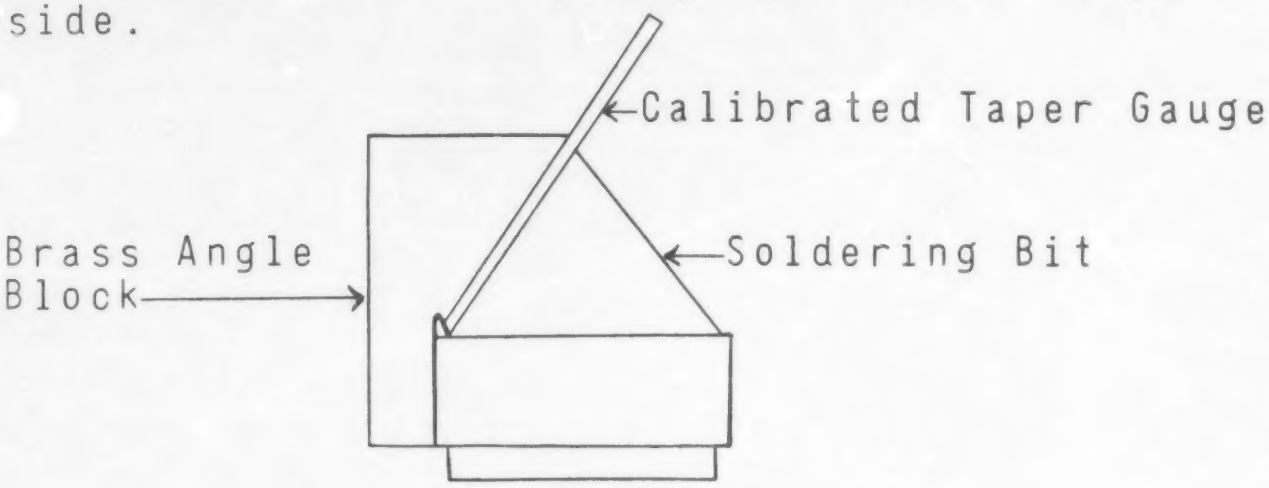
2 Check and square 90° apart 2½ marks each way.

SURFACE FINISH

Visual marking for general finish.

ANGLES

Using both No. 2 Gauges (Brass angle block and Calibrated taper gauge). Fit taper gauge between Solder Bit and Angle Block. Read off mark on taper gauge. 5 marks per side.



HANDLE CENTRE

Visual check using Gauge (No. 3).